

Role of Natural Language Processing and Machine Learning on (MTS)-Machine Translation System for Indian Languages: A Review

Megha V¹ and Pooja M R²

¹⁻²Vidyavardhaka College of Engineering, Mysuru, India

Email: megha.v@vvce.ac.in, pooja.mr@vvce.ac.in

Abstract—The conversion process of one specific language to another either in a completely automatic manner or with considerable amount of human intervention by preserving the original meaning of the input source text is known as (MTS) Machine Translation System. The expeditious availability of data and the need for increased processing speed of information has led the way towards Natural Language Processing, Machine Learning, Speech Processing, Image Processing etc. When it comes to the Indian regional languages the research in this domain is very minimal as an outcome there is limited resources. In this paper, the researcher will be able to know the general workflow of how the translation process of source to destination language will occur and the various approaches and the methods for processing Indian languages proposed by scientists which are verified and validated, the challenges a researcher would face while developing one such translation tool followed by the summary of the entire review.

Index Terms—Machine Translation System; Machine Learning; Natural Language Processing; Translation model; Rule-based Machine Translation; Statistical Based Machine Translation; Neural Machine Translation; Hybrid Machine Translation.

I. INTRODUCTION

Any spoken language that has developed with one's own accord is known as original language, it allows people to exchange their knowledge, sentiment, and emotions with individuals. The native language has its unique syntax, signs, and alphabets. The sole nation is India with rich diverse variations of regional languages. To translate these diverse languages into any other language needs a good machine translation system [16]. Human-assisted translation of one language into another is known as traditional machine translation wherein human intervention is required to translate the language with best interpretation, the major drawback in this system is that time commitment is required from the humans who are involved in the translation process. Some evolution with respect to the translation process was required to make the translation process quick and accurate, the expense of doing human assessments of MTS output is sometimes prohibitive when evaluation needs to be done quickly and often to monitor progress.

A long-term objectives of computer science is machine translation for natural languages, which is the sub-domain of artificial intelligence and machine learning where natural language processing (NLP), Computational Language studies them in PC-based systems. Processing, interpretation, and creation are the three essential

components of NLP [3]

An artificial intelligence subfield (AI) is called "natural language processing" (NLP) enables computers to understand, produce, and manipulate human language. Data can be questioned using natural language text or voice thanks to natural language processing. Another name for this is "language in." Without realizing it, most consumers have undoubtedly engaged with NLP. For instance, the Oracle Digital Assistant (ODA), Siri, Cortana, or Alexa are all virtual assistants that use NLP as their primary technology. NLP is what gives these virtual assistants the ability to answer in natural language to our enquiries while also comprehending our requests. NLP is applicable to both spoken and written communication and to all human languages. Web search, spam screening for emails, automatic translation of text or speech, document summarization, sentiment analysis, and grammar/spell checking are more examples of technologies that use NLP. NLP is used by these applications to read, analyse, and respond to your communication. For instance, certain email programs can automatically recommend a suitable reply to a message based on its content.

The difficulty of machine translation in the Indian context is far more complex. Morphological complexity and diversity are two important factors that make parallel translation—where the corpus is restricted to different languages—even more difficult [17]. Traditional machine translation algorithms were developed using Rule-based and Statistical techniques to overcome this complexity, but they were unable to provide the correct translation mechanism.

II. MOTIVATION

The phrase "UNITY IN DIVERSITY" is used to characterize a nation such as India, which has 122 major languages spoken across its 29 states and 7 union territories, each of which has its own significance and ties to the local community. Every researcher will have a strong connection to their home language [2], which opens the door for problem statements pertaining to the translation system from any preferred source language to their native tongue so that people would benefit more. The concept of machine learning was first proposed in 1933. The USSR Academy of Sciences received the concept from Soviet scientist Peter. Twenty years passed before his innovation was recognized, and two Soviet.

III. CATEGORIES OF MACHINE TRANSLATION/ MACHINE TRANSLATION TECHNIQUES

Traditional machine translation and improvised machine translation systems are the two main categories into which the MTS i.e, machine translation system can be roughly divided.

A. Traditional Machine Translation

As mentioned in the introduction, any translation with human intervention i.e. human translator who has got awareness of the destination and source languages by understating certain cultural elements that, while perhaps existing in the source text, cannot be translated into the language of the destination text using an equivalent phrase, as automatic translation does, to give the result of the translated text while maintaining the original source text's meaning. Among the most difficult obligations in the translating process is manual understanding.



Fig 1. Representation of Traditional MTS

B. Machine Translation System (MTS)

Along with the current system, several machine translation programs are in use like i) Rule-Based Machine Translation ii) Statistical Machine Translation iii) Neural Machine Translation iv) Hybrid Machine Translation

Rule-based Machine Translation:

Major deconstruction in this translation is that it involves an analyst who analyses the original source text and gives the output structure for the source text given. Targeted destination language pairing consists of million words and are typically bilingual, monolingual, or multilingual dictionaries. The integration of word-based approach i.e. RBMT and the phrase-based systems enhance the accuracy of the translation and the user feasibility.

Statistical Machine Translation or Statistical Based Machine Translation (SBMT/SMT):

The translation constructed using statistical language models from a large text sample. This a strategy is utilized with the integration of a corpus and is very much effective in determining which models are needed for translation.

Neural Machine Translation (NMT)

It is made up of neurons which work as a node to create a grid pattern, this network is made up of decoders and encoders that are utilised in a layered architecture to provide traditional output. Neural machine translation, or NMT for short, is the use of neural network models to learn a statistical model for machine translation. The key benefit to the approach is that a single system can be trained directly on source and target text, no longer requiring the pipeline of specialized systems used in statistical machine learning. The SBMT systems were modified with the discovery of NMT. Neural machine translation which in turn increases the better outcome and the fluency of the translation.

Hybrid Machine Translation (HMT):

It is a type of MT which inherits the working of multiple translation systems, single machine translation systems failed to give superior outcomes in terms of accuracy and lead to the development of Hybrid Machine Translation. Machine translation systems like Omniscien Technologies and PROMP SYSTRAN, are the hybrid mode of machine translation systems.

IV. THE GENERAL MODEL OF NATURAL LANGUAGE PROCESSING

The general model of natural language processing includes various stages like machine transliteration, preprocessing, lexical analysis and POS tagging, morphological analysis, feature extraction and evaluation.

A. Tokenization

The method of transforming raw data into meaningful data string is known as tokenization, where a raw data set once given as a input to the tokenizer will give the tokens which is the most important fundamental lexical unit. Based on the problem statement, tokenization may take place at the word level or at the phrase level. Sentence level tokenization is one among the methods of tokenization, Word-level tokenization, N-gram tokenization. The identification of end of sentence and sentence boundary disambiguation is the primary emphasis of sentence level tokenization. Word level tokenization is the most popular tokenization technique since it treats words as lexical units and divides the entire document into a set of words. The n-gram tokenization is a token made from n words, where 'n' denotes the number of words that are taken into consideration as a single lexical unit. If n=1, bigram if n=2, or trigram if n=3, the model is referred to as a unigram model.

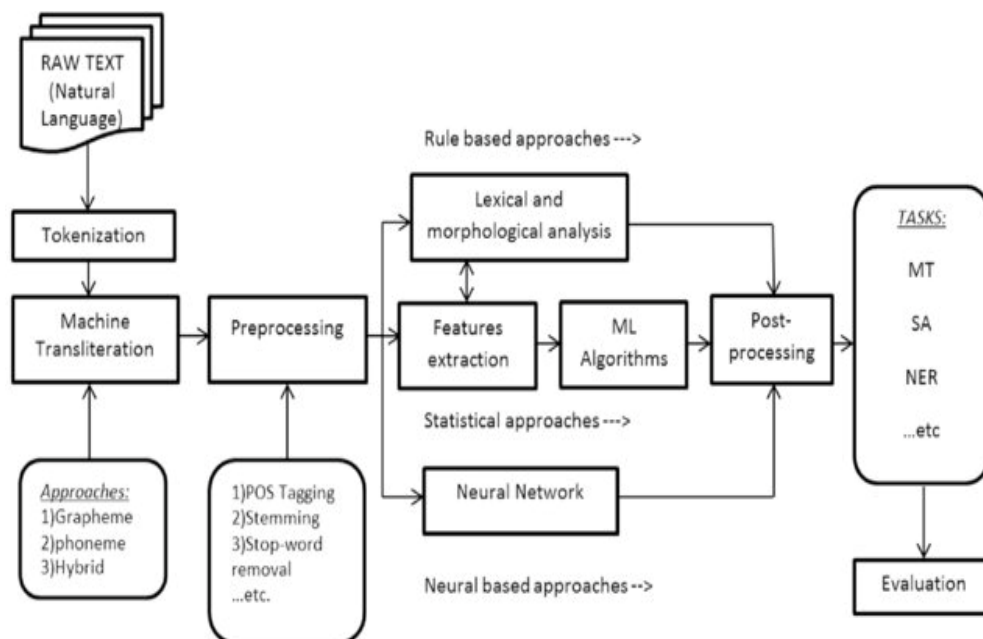


Fig 2: Generic Model for Natural Language Processing [3]

B. Machine Transliteration

Machine transliteration, named entity identification, information retrieval, and other related applications are significantly reliant on cross-lingual machine translation. Without affecting the phonetic of the source language, the source text is transformed into the destination language's alphabetic system. There are three categories of machine translation. Phoneme, grapheme, and hybrid [12] Without a consideration the source language's phonetic knowledge, the grapheme-based transliteration model directly transliterates the source language words into the target language words. Phoneme-based transliteration employs the phonetics of the source language word to convert it to the grapheme of the target language, as suggested by its name. The hybrid transliteration process converts the source language's phoneme and grapheme into the target language's grapheme.

C. Preprocessing Techniques

Tokenization and transliteration steps yield the tokens and the phonetics of the raw text once these actions are finished preprocessing technique is used to increase the applications effectiveness, as per the requirement. Major techniques like stemming, stop-word removal, Lemmatization, POS tagging, Unicode Normalization.

Stemming

Normalization of the stated phrase into its stem/root i.e. identification of the root word and to give its grammatical categories which can occur in a sentence, which is called as the inflection of the language. Numerous stemming algorithms exist for Indian languages and non-Indian language.

Stop-word removal

Removal of repetitive words which is insignificant towards adding on any meaning to the sentence is called as stop-word removal this preprocessing technique is required in order to make things easier of language processing, techniques like Deterministic Finite Automata based stop-word removal, dictionary based-stop words removal are widely being used. [18]

Lemmatization

Lemmatization is a method that reduces a word's prefix and suffix to a form that is acceptable in the target language; the resulting term is referred to as a "Lemma." For instance, in the English language, the word "draw" can take a variety of forms, such as "draws," "drawing," and "drawed." To which the result is, "draw" is the lemma for all those forms.

POS Tagging

The outcome of grammatical differences between languages, the procedure by which the parts of speech (POS) such as Noun, Pronoun, Verb, Adjective, etc. will be appended to the lemma is referred to as POS tagging. In the realm of NLP, POS tagging aids in sentiment analysis, named entity recognition, cross-lingual machine translation, and any research pertaining to POS tagging aids in the development of several native language applications.

Unicode Normalization

The worldwide character encoding system known as Unicode assigns each text character a distinct hexadecimal value. Unicode normalization converts equivalent sequences of characters into the same representation, which is utilized for information processing.

D. Post-preprocessing

Post-preprocessing step includes the verification of the outcomes from the preprocessing steps and the results are made much more refined or efficient. Validation is completed with respect to spelling mistakes, sentence arrangements, missed translations, grammatical errors etc., and any fine tuning required for the increase in efficiency for the results will be performed.

E. Evaluation

For machine translation projects, BLEU scores are generated to evaluate the translation's quality; it is a quantitative metric for assessing the calibre of an output text based on a few reference texts. When measuring OOV words in translation activities, the Unknown Word (UNK) count is used, while the Word Error Rate (WER) measure is used to examine both human and machine translation output.

V. CHALLENGES

Natural language processing is undoubtedly a dynamic tool with huge benefits, on the contrary there are considerable number of obstacles encountered in creating such effective tool which can be summarized for Indian Regional Language are as follows:

A. Contextual words and phrases and homonyms

Depending on the context, only a single word might have a variety of interpretations, for example in the English language we have sentences like “*I ran to the store because we ran out of milk*”, “*The house is looking really run down*”, the word ran in the first sentence has got two meanings with regards to its context, in the second sentence the word run has got a several meanings which is way far from the actual meaning with regards to the context which is mentioned in the sentence, It is really simple for anybody with the English knowledge to understand these context, despite the fact that NLP language models may have learned every meaning, distinguishing between them in context might be challenging. Homonyms, which consist of just two or more similar-sounding words but mean different things, can be challenging for speech-to-text and question-answering applications because they aren't written in text. Even for humans, misuse of the words there and their is a prevalent issue.

B. Synonyms, Irony and Sarcasm

Since we employ a variety of words to convey the same idea, synonyms might present problems akin to those associated with contextual awareness. Additionally, while some of these terms may have the exact same meaning, others may be more complicated (such as little, little, tiny, and minute), and different people may use synonyms to indicate somewhat different meanings in their own vocabularies. As a result, it's crucial to include all conceivable synonyms and definitions of words while developing NLP systems. While text analysis algorithms may still make errors now and then, their comprehension of synonyms will improve as they get more pertinent training data. Since user frequently use words and phrases that, precisely by definition, may be positive or negative but actually evince the contrary, irony and sarcasm are the current problems faced in machine learning models. Models can be taught using word embeddings and certain cues that regularly appear with sardonic or sarcastic phrases, such as "yeah right," "whatever," etc., but it's still a challenging process to train the machine appropriately.

C. Ambiguity

Ambiguity in NLP refers to the interpretation faults which can happen while translating this usually happen when a single word has multiple meanings with respect to the context. *Lexical Ambiguity*: A word that could be used as a verb, noun, or adjective. *Semantic Ambiguity*: The interpretation of a sentence in context of usage. *Syntactic Ambiguity*: For example: I saw the boy on the beach with my binoculars. This could mean that I saw a boy through my binoculars or the boy had my binoculars with him. This is the syntactic ambiguity which can be encountered.

D. Errors in text or Speech, Colloquialisms and slangs and Domain-Specific Language

Words that have been misused or misspelled can cause issues with text analysis. However, these problems can be reduced as language databases expand and intelligent assistants are trained by their unique users. For NLP, informal expressions, idioms, and cultural jargon create a number of challenges, particularly for models aimed at a broad audience. There may be no "dictionary definition" at all for colloquialisms, and these terms may even have various regional meanings. Additionally, since cultural slang evolves and grows, new terms appear every day. This is where Models of any type can be trained and updated often, albeit this frequently necessitates a large amount of data. It's possible that different businesses and industries utilize very different languages. For instance, the NLP processing model needed to process medical data might be very different from the one needed to handle legal documents.

E. Low-Resource Language and Lack of research and development

However, many languages, particularly those used by those who have less access to technology, are frequently underutilized and ignored. To the fullest extent possible, machine learning needs BILLIONS of pieces of training data. NLP models get smarter the more data they are trained on. To increase efficiency, more research must be done on novel approaches to this kind of problem.

VI. SUMMARY

As per the survey findings 72.1% of online customers spend their time on sites in their native language and 72% of customers prefer to buy products with the information in their mother tongue. It is a common observation that any communication would be effective if the mode of communication is done through the native language or mother tongue. These are the elementary reasons that it is essential to train machines to learn various languages and communicate with users in their desired language. This paper has summarized various ways in which such machine translation would happen and the processing techniques it undergoes along with the challenges one could face while designing such NLP systems. There is still a need for good quality dictionaries such as WordNets, Corpora for the less resourced Indian languages. There are pros and cons with respect to each approaches for the translation system, i.e rule based approach is inadequate while making complex rules and while processing the languages which do not have proper corpora, it is one of the approach where in there is supposed to be a regular updation with respect to the rules which are being framed for any minute evolution in the language dictionary. In the cases where contextual based analysis is required i.e to do the sentiment analysis of the sentence considered rule-based approaches are feasible, to quote about the significance of statistical based approaches it does not require linguistic acquaintances. To the languages in which the corpus is less, the statistical based approach finds the best fit, Neural based approach provides the best solutions to various tasks, a major drawback of Neural based approach is that if the corpus is less then obtaining the desired result will be very difficult. Analysing the proper fit of approaches and the algorithm along with the good number of corpus for the problem statement defined also plays an important role in obtaining the accurate translation.

REFERENCES

- [1] Pushpalatha Kadavigere Nagaraj, Kshamitha Shoba Ravikumar, Mydugolam Sreenivas Kasyap, Medhini Hullumakki Srinivas Murthy, Jithin Paul Kannada to English Machine Translation Using Deep Neural Network. In:
- [2] Gunti Spandan, Dr. Brahmananda S H (2023) Impact of Machine Learning on Regional Languages Processing: A Survey. In: 2nd International Conference for Innovation in Technology (INOCON) Bangalore, India. Mar 3-5, 2023
- [3] B. S Harish, R Kasturi Rangan A Comprehensive survey on Indian regional language processing for SN Applied Sciences Springer Nature Switzerland AG (2020)
- [4] Aarati H Patil, Snehal S Patil, Shubham M Patil, and Tatwadharshi P Nagarhalli. Real time machine translation system between indian languages. In 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), pages 1778–1783. IEEE, (2022)
- [5] Tatwadharshi P Nagarhalli, Vinod Vaze, and NK Rana. A novel framework for neural machine translation of indian-english languages. In 2020 International Conference on Inventive Computation Technologies (ICICT), pages 676–682. IEEE, (2020)
- [6] Maheshi Nandathilaka, Supunmali Ahangama, and G. Thilini Weerasuriya. A rule-based lemmatizing approach for sinhala language. In 2018 3rd International Conference on Information Technology Research (ICITR), pages 1–5, (2018)
- [7] Sandeep Saini and Vineet Sahula. Neural machine translation for English to hindi. In 2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP), pages 1–6. IEEE, (2018)
- [8] Ravi Narayan, S Chakraverty, and VP Singh. Neural network based parts of speech tagger for hindi. IFAC Proceedings Volumes, 47(1):519–524, (2014)
- [9] Amarappa S, Sathyanarayana S Named entity recognition and classification in kannada language. Int J Electron Comput Sci Eng 2(1):281–289 (2013)
- [10] Antony P, Soman K Kernel based part of speech tagger for kannada. In: 2010 international conference on machine learning and cybernetics, IEEE, vol 4, pp 2139–2144 (2010)
- [11] Antony P, Soman K (2011) Machine transliteration for indian languages: a literature survey. Int J Sci Eng Res IJSER 2:1–8
- [12] Antony P, Ajith V, Soman K (2010) Kernel method for english to kannada transliteration. In: 2010 international conference on recent trends in information. Telecommunication and Computing, IEEE, pp 336–338
- [13] Bhargava R, Arora S, Sharma Y (2019) Neural network-based architecture for sentiment analysis in Indian languages. J Intell Syst 28(3):361–375
- [14] Godase A, Govilkar S (2015) Machine translation development for Indian languages and its approaches. Int J Nat Lang Comput 4:55–74
- [15] Kaur B, Veer D (2016) Translation challenges and universal networking language. Int J Comput Appl 133(15):36–40
- [16] Seema Bawa, Munish Kumar, et al. A comprehensive survey on machine translation for english, hindi and sanskrit languages. Journal of Ambient Intelligence and Humanized Computing, pages 1–34, 2021.
- [17] Sandeep Saini and Vineet Sahula. A survey of machine translation techniques and systems for indian languages. In 2015 IEEE International Conference on Computational Intelligence & Communication Technology, pages 676–681. IEEE, 2015.

- [18] Jha V, Manjunath N, Shenoy PD, Venugopal K (2016) Hsra: Hindi stopword removal algorithm. In: 2016 international conference on microelectronics. Computing and Communications (Micro-Com), IEEE, pp 1–5
- [19] Middi Venkata Sai Rishita¹, Middi Appala Raju, and Tanvir Ahmed Harris ¹RNS Institute of Technology, India²Christ University, India , IIT Bombay India Machine translation using natural language processing for MATEC Web of Conferences 277, 02004 (2019)
- [20] Padma M.C. and Prathibha R.J., “Development of morphological stemmer, analyzer and generator for Kannada nouns”, International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), published in Lecture Notes in Electrical Engineering (LNEE) Series, Springer India, ISSN:1876-1100, pp. 713-723, 2014.
- [21] Sindhu Satish Kumar Prathibha R. J., Sharanya Kumari Shivakumar Recurrent Neural Network Based Parts of Speech Tagger For Kannada Words For International Conference On Recent Trends In Science & Technology (ICRTST - 2021)